DOCUMENT RESUME

ED 141 384 TM 006 272

AUTHOR Hsu, Yi-Ming; Scott, Owen
TITLE An Inventory for Appraising Experimental Research Designed for Introductory Research Methods Classes.
PUB DATE Apr 77
NOTE 44p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS College Students; *Educational Research; *Evaluation Criteria; Higher Education; *Psychological Studies; *Rating Scales; Research Methodology; *Technical Reports; Test Construction; *Test Reliability; Test Validity
IDENTIFIERS *Evaluation Instrument for Experimental Research

ABSTRACT

The development of the Evaluation Instrument for Experimental Research (EIFER), an inventory for appraising research quality, is described. The EIFER was specifically designed to aid students in introductory courses in educational and psychological research methods to evaluate published research reports. A survey conducted among specialists in educational research resulted in a list of 73 characteristics considered essential in experimental and quasi-experimental research. These characteristics were organized into 6 categories: the research problem, review of related literature, research design, data collection and analysis, conclusions and generalizations, and style and organization. The following psychometric properties of the inventory were determined: consistency, item-section correlation, analysis of variance, interrater and intrarater reliability, and stability of the measurement of traits across raters and occasions. Reliability coefficients were determined to be satisfactory. A copy of the instrument is included. (Author/GDC)

An Inventory for Appraising Experimental Research Designed for Introductory Research Methods Classes

Yi-Ming Hsu
West Chester State College

Owen Scott
The University of Georgia

Paper presented at the Annual Meeting of the American Educational Research Association, New York City, April, 1977.

An Inventory for Appraising Experimental Research, Designed for Introductory Research Methods Classes[1]

Yi-Ming Hsu
West Chester State College

Owen Scott
The University of Georgia

Introduction

Research studies vary widely in quality and published reports of them differ significantly in value. Therefore, an answer to the question, "How much confidence can be placed in the results reported in a study?" is one of great importance. Although professional journals of education devote considerable space to reports of educational research, many readers have difficulty in evaluating their worth. One reason for this difficulty may be the absence of agreed-upon criteria in terms of which to make the judgments which are simple enough for a statistically naive reader to apply. Although there have been many proposals for appraising published research reports in education and psychology (Brooks, 1923; Davitz & Davitz, 1967; Farquhar & Krumboltz, 1959; Fox, 1958; Ingle & Gephart, 1966; Suydam, 1968, Wandt, 1965; Ward, Hall, & Schramm, 1975), almost all of them, however, have been quite complex and demand considerable sophistication on the part of their users.

It was the intent of the research reported here to produce a set of criteria which can be used to evaluate important

[1] Based on the unpublished doctoral dissertation of the first author.

aspects of published research reports and yet which are simple enough to be understood and applied by neophytes who have had guided training in their use in beginning courses in methods of educational research. The development of these criteria, then, rests upon these premises: (1) essential characteristics of valid research can be identified; (2) some of these are simple enough to be understood and applied by students with limited backgrounds who are enrolled in a first course in educational research methods; and (3) if these students study the established criteria carefully and have guided practice in the appraisal of research in terms of them, they should become more proficient in making sound judgments concerning the internal as well as external validity of the results reported in a specific study.

Purpose

The objective of the study was to develop an inventory for appraising experimental and quasi-experimental research designed specifically for introductory research methods classes in education and psychology. The inventory should posses such essential characteristics as content validity, internal consistency, and stability-of-response reliabilities.

An extensive review of literature had revealed that quite a few of appraisal forms had been developed for the use of evaluating reported educational research (Hsu, 1975). They would generally appear in one of the three forms: article-type, checklists, and rating instruments. Regardless of the format

in which the instruments were presented, they tended to contain one or more of the flaws identified below:

1. The procedures of development were not described.
2. A key or scales were not provided for judging the criteria included.
3. The specifications of the scale, if available, were either undefined or very ambiguous.
4. No empirical data were generated to support the adequacy of the key or scales provided.
5. Psychometric characteristics of the instrument were either incomplete or not established at all.
6. The type of research for which the particular instrument is intended was not specified.

As a matter of fact, none of the evaluation instruments identified by the authors possessed the three attributes which characterize the inventory developed in this study: (1) it is intended for use in teaching the beginning students of research methods the key points in the appraisal of experimental educational research; (2) such essential psychometric information as the inter-rater and intra-rater reliability indices was checked and established; and (3) its appropriate usefulness for instructional purposes was explored empirically through an experiment conducted in classes of research methods at the University of Georgia, Athens, Georgia.

Procedures

The overall objective of this study was to develop an inventory of essential characteristics of experimental research. In order to accomplish this task, several stages were formulated as shown in Figure 1.

- - - - - - - - - - - - - - - -

Insert Figure 1 about here

- - - - - - - - - - - - - - - -

Establishment of Content Validity

To ascertain the suitability of the evaluation instrument as an instructional aid to students of the beginning research methods classes, its content validity must be properly established. The following steps were taken to achieve this goal:

1. A survey was made of texts on research methods and of journal literature reporting the construction and use of instruments for appraising educational research.
2. Based on the survey, a list was prepared of characteristics essential to the execution and reporting of experimental and quasi-experimental research.
3. The list was then sent to a nation-wide panel of 35 educational research specialists for their independent judgments as to the essentiality of each characteristic included.
4. Characteristic judged as essential by 17 or more of the 21 specialists responding were retained. Based on this

criterion, 46 out of the 51 items identified in Step 2 above were retained in their original forms or retained after revision.

5. The list of retained characteristics was again given a nation-wide panel of 40 educational research methods instructors listed in the membership directory of the American Educational Research Association Special Interest Group: Professors of Educational Research. The panel was asked to independently appraise each characteristic as one appropriate for inclusion in introductory research methods courses.
6. Characteristics judged appropriate by 30 or more of the 34 who responded were included in the final version of the inventory. In addition, comments or suggestions given by the members of the panel were used as guidelines for rearranging or restructuring some of the statements, if deemed necessary.

Selection of Foils

A set of structured responses (foils) was needed for use with each inventory statement of an essential characteristic, thereby enabling the student to appraise the research in terms of that characteristic more precisely by selecting one of the foils. A review of similar inventories and logical considerations reduced the types of options considered to only two: categorical versus continuous. The categorical type was equipped with a definition for each of the five responses

listed while the continuous type had a five-point scale, ranging from 1 for "inadequate" to 5 for "adequate", with 2, 3, and 4 set in between undefined.

For the selection of the key to be incorporated in the inventory, a pilot study was conducted in a graduate class of an advanced educational psychology course. The students were randomly divided into two groups, with one using the categorical key and the other the continuous type, to rate an article of experimental research in education. On the basis of the empirical comparison made in the pilot study, one set of the foils was selected for use with the evaluation instrument.

Check on Internal Consistency and Stability-of-Response Reliabilities

Since the inventory was developed primarily for the evaluation of published reports of research in experimental or quasi-experimental design, a sample of research reports of this nature was selected for use in the pilot study and in a subsequent study to check on the psychometric properties of the instrument. Articles in the American Educational Research Journal (AERJ) published in the most recent four years of 1970-73, inclusive, were surveyed to identify research articles of this specific design. Two were randomly selected, approved for use by AERJ, and "blinded" as to author and publication. Eight instructors with experience in teaching graduate level research methods at the University of Georgia were asked to use the inventory and appraise the two selected articles. Without foreknowledge of the request, each was asked to reappraise the two articles

approximately one month later. These appraisals provided data for Raters by Articles by Traits by Occations analyses of variance (ANOVAs) and a portion of the data for the Kuder-Richardson Formula 20 (K-R 20) reliability estimates (see Figure 2 for the lay-out of these ANOVAs and reliability estimates). Moreover, members of a graduate research seminar

- - - - - - - - - - - - - - - -

Insert Figure 2 about here

- - - - - - - - - - - - - - - -

also appraised and re-appraised the two articles at about the same time and under the same instructions as the instructors. Eleven of the 13 graduate students completed two evaluations of each articles. Data from all of these sets of appraisals were used for a series of estimates of the K-R 20 reliability coefficients.

Check on the Usefulness of the Inventory

The attributes contained in the inventory were essential characteristics of experimental research. They were prepared specifically for use in educating beginning students of research methods. Hence it was necessary to ascertain its usefulness by conducting an experiment in classes of research methods.

In this experiment, the experimental group (E) consisted of a random half of each of five introductory research methods classes. They (with a total of 58 Ss) were given copies of the inventory and of one of the two articles. The other half from the same five classes constituted the control group (C)

with a total of 55 subjects. They were given copies of the same article but not the inventories. Both groups were asked to appraise each of the six "global" aspects (corresponding to the six inventory sections) of the article. The hypothesis tested was that the proportions of inventory-users whose "global" appraisals agreed with those of the eight instructors would be greater than the proportions of non-users who agreed. The data were obtained a few days before the end of the quarter.

Results

Inventory Content

In its final form the inventory, Evaluation Instrument for Experimental Research (EIFER), contains a list of 73 characteristics categorized into six major sections, each pertaining to a major aspect of experimental research - Research Problem, Related Literature, Research Design, Data Collection and Analysis, Conclusions and Generalizations, and Style and Organization of the Report. At the end of each section is an item pertaining to a "global" appraisal with respect to that particular section (see Appendix 1).

Structure of the Foils

From the results of the pilot study, an empirical comparison was made of the two sets of the foils mentioned earlier on the basis of the variabilities of the appraisals with each set, which were determined from the sum of weighted absolute deviations from the modal response, as well as of a

survey of user's preference for the keys. The empirical findings strongly favored the same set, the categorical type, which is listed below (also see Appendix 1):

1: The article contains no information concerning the attribute.

2: The information given clearly indicates that the attribute was inappropriately handled.

3: The information given suggests that the attribute may have been improperly managed.

4: The information given indicates that the attribute was properly managed.

5: The information given clearly indicates that the attribute was appropriately handled.

Estimates of Selected Psychometric Characteristics

The primary psychometric measures estimated for EIFER included: measures of internal consistency in terms of KR-20 indices, item-section intercorrelations, analysis of variance (ANOVA) for each section of the inventory, interrater reliabilities, intrarater reliabilities, and stabilities of measures of traits across raters and occasions.

Measures of internal consistency. To test the homogeneity of inventory items, an estimate of reliability was produced via KR-20 (r_{tt}). These reliability coefficients are the average correlations obtained from all possible split-half reliability estimates. As shown in Table 1, all nine estimated coefficients for total inventory score were .90 or above, six of the nine

- - - - - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - - - - -

for the set of "global" appraisals were .80 or above, and of the 54 section KR-20s, 17 of the 33 with .80 or higher were .90 or above while only four were lower than .60.

Table 2 contains the correlations between EIFER sections.

- - - - - - - - - - - - - - - -

Insert Table 2 about here

- - - - - - - - - - - - - - - -

The correlation coefficients ranged from the lowest ($r = .28$) between Research Problem and Conclusions and Generalizations to the highest ($r = .69$) between Research Design and Data Collection and Analysis, with most of the pairwise relationship of other sections falling within the moderate range. This gave a rather clear indication that the attributes contained in each section were not repeated of nor overlapped with those across other sections. Such an indication, in turn, supported the accomplishment of the major function by each EIFER section in measuring different rather than similar or identical aspect of the reported research article.

Item-section intercorrelations. For the purpose of checking on the "goodness of fit" of each inventory item in the particular section, a matrix of item-section correlations was generated (see Table 3). Ideally, a specific item should correlate higher

- - - - - - - - - - - - - - - -

Insert Table 3 about here

- - - - - - - - - - - - - - - -

with its own section than with others of the inventory to validate its "legitimate nesting" in that particular section. Evidently, except for some of the items in the Related Literature section, a great majority of the items in the other six EIFER sections correlated higher with the section in which they were included than with others of the inventory.

Analysis of variance. Using the appraisal and reappraisal of the two selected articles by the eight instructors of educational research methods, an ANOVA was performed for each of the six EIFER sections. The primary purpose of the analyses was to obtain reliability data from the significance tests for main and interaction effects of the four factors involved in the study (i.e., "Article", "Rater", "Trait", and "Occasion") and thereby to estimate reliability coefficients on the basis of the various variance components.

In all of the ANOVA's, "Occasion" and "Trait" were considered to be fixed factors; "Rater" and "Article", random. In theory, however, the appropriate error term is not available for the source of variance of a fixed effect variable. Accordingly, Myers (1972) suggests applying a quasi-F ratio to test the statistical significance of such an effect. In this study, Myers' technique was followed to test the main effect of both "Trait" and "Occasion" as well as their associated interaction effects.

Table 4 shows the results of ANOVA for the section of Research Problem. The main effect of "Trait" was found highly

- - - - - - - - - - - - - - -

Insert Table 4 about here

- - - - - - - - - - - - - - -

significant at the .01 level (F (12, 11.81) = 6.56, $p < .01$). A statistically significant "Trait" main effect indicates that appraisals, averaged across raters, articles, and occasions differed from trait to trait to some extent, obviously, a desirable attribute for the inventory to possess.

In the analysis for the section of Related Literature (see Table 5), the main effect of "Article" was found significant

- - - - - - - - - - - - - - -

Insert Table 5 about here

- - - - - - - - - - - - - - -

(F (1, 7) = 10.23, $p < .05$). So was the interaction effect between "Article" and "Trait" (F (5,35) = 2.90, $p < .05$). Both outcomes were desirable and favorable to this section of EIFER in the sense that the appraisals differentiated article from article on the various traits presented in the instrument, across raters and occasions as well.

The analysis performed on the section of Research Design resulted in both desirable and undesirable effects. As shown in Table 6, the desirable main effect of "Trait" was highly

- - - - - - - - - - - - - - -

Insert Table 6 about here

- - - - - - - - - - - - - - -

significant at the .001 level (F (29, 43.13) = 5.35, $p < .001$). The undesirable outcomes were mainly due to the various interaction effects.

Table 7 presents the results of analysis for the section of Data Collection and Analysis. In addition to the highly

- - - - - - - - - - - - - - -

Insert Table 7 about here

- - - - - - - - - - - - - - -

significant main effect of "Article", two interaction effects were found significant, one desirable, between "Article" and "Trait" and the other undesirable, between "Rater" and "Trait".

Two slightly different pictures were drawn from the analyses of the last two sections of the inventory. As shown in Table 8, the "Trait" main effect was found statistically

- - - - - - - - - - - - - - -

Insert Table 8 about here

- - - - - - - - - - - - - - -

significant in the section of Conclusions and Generalizations (F (9, 11.12) = 5.20, $p < .01$), but not in that of Style and Organization as presented in Table 9. On the contrary, a significant "Article" main effect (F (1, 7) = 6.22, $p < .05$) was found in the analysis for the section of Style and Organization (see Table 9), but not in that for Conclusion and Generalizations.

- - - - - - - - - - - - - - -

Insert Table 9 about here

- - - - - - - - - - - - - - -

When the entire inventory items were combined and analyzed, excluding the 6 overall evaluation items, two desirable main effects were found significant, i.e., "Article" and "Trait" (see Table 10), in addition to one first order interaction and

- - - - - - - - - - - - - - - -

Insert Table 10 about here

- - - - - - - - - - - - - - - -

two second order interactions. On the basis of the findings from the various analyses performed, it is quite evident that as a whole the inventory generated satisfactory internal consistency and stability-of-response reliabilities.

Estimates of reliability indices. The results of the ANOVAs were used to estimate three kinds of reliability coefficients (i.e., interrater, intrarater, and stability of measure of traits) publicized by Stanley and Wiley (1962) and rather easily estimated by procedures developed and described by Silverstein (1974).

The interrater reliability coefficient is the average correlation between the ratings of an inventory item made by the eight judges, with the average obtained across articles and occasions. The intrarater reliability coefficient is the average correlation between a judge's appraisals and reappraisals of the items in an inventory section averaged across raters and occasions. The stability of trait measurement reliability coefficient is the average correlation between the judges' appraisals and reappraisals of an inventory item averaged across

articles and traits. Table 11 contains the estimated indices

- - - - - - - - - - - - - - - -

Insert Table 11 about here

- - - - - - - - - - - - - - - -

for the three types of reliabilities. The estimates were computed for each of the six inventory sections, the EIFER composite, and the "global" evaluations pertaining to the six major aspects of experimental research.

Results of the "usefulness" experiment. For each of the six EIFER sections, the proportion of users in agreement with the "global" appraisals of the eight instructors was compared with the proportion of non-users. As shown in Table 12, the

- - - - - - - - - - - - - - - -

Insert Table 12 about here

- - - - - - - - - - - - - - - -

smallest of the 12 proportions was .72, with the other 11 in the interval, .76 - .89. None of the differences was statistically significant at the .05 significance level.

Two circumstances may account for the high proportions of agreement and for the non-statistically significant differences. The comparisons were made near the end of the quarter after both groups had had identical experiences in research appraisal. Moreover, some of the instructors had stressed many of the characteristics contained in EIFER. For reasons pointed out in the report of which this is a summary, the experiment as conducted differed in important respects from the experiment as originally planned.

Conclusions

The goal of improving the quality of educational research is not new among professionals in educational circles. Such a goal will be far from reaching or is likely to be illusory, however, unless professional concern is translated into serious efforts. One such effort is to develop an evaluation instrument to assist the prospective educational researcher to appraise published experimental research studies critically and yet objectively. EIFER was developed especially for such a purpose.

On the basis of the empirical evidence obtained in this study, the inventory has been characterized with the satisfactory psychometric properties deemed to be essential, though some improvement is desirable. First of all, the results of item analysis gave a clear indication that most inventory items were adequately nested in the sections to which they should belong. Then, the structure of both separate sections and the complete inventory was effective and favorable as explored from various reliability estimates via K-R 20 approach. Furthermore, the desirable outcomes of main effects from a series of ANOVA tests demonstrated that use of the inventory would result in differential appraisal of the two different research reports with separate characteristics, as proved with satisfactory internal consistency and stability-of-response reliabilities.

It is reasonable to conclude that EIFER in its present form is a satisfactory aid to teaching students of research methods classes. Since the essential characteristics of

experimental research are basically the same regardless of the discipline of interest, the inventory should prove useful not only in introductory research methods classes in education and psychology but in other research methods classes within the spectacular domain of social sciences.

References

Brooks, F. D. Criteria of educational research. School and Society, 1923, 18, 724-729.

Davitz, J. R., & Davitz, L. J. A guide for evaluating research plans in psychology and education. New York: Teachers College Press, 1967.

Farquhar, W. W., & Krumboltz, J. D. A checklist for evaluating experimental research in psychology and education. Journal of Educational Research, 1959, 52, 353-354.

Fox, J. H. Criteria of good research. Phi Delta Kappan, 1958, 39, 284-286.

Hsu, Y. Development and validation of an instrument for evaluating experimental educational research in research methods classes. (Doctoral dissertation, the University of Georgia) Ann Arbor, Mich.: University Microfilms, 1976, No. 76-6409.

Ingle, R. B., & Gephart, W. J. A critique of a research report: programmed instruction versus usual classroom procedures in teaching boys to read. American Educational Research Journal, 1966, 3, 49-53.

Myers, J. L. Fundamentals of experimental design (2nd ed.). Boston, Mass.: Allyn and Bacon, 1972.

Silverstein, A. B. Interrelationships between analysis of variance and correlational analysis. Educational and Psychological Measurement, 1974, 34, 801-805.

Stanley, J. C., & Wiley, D. E. Development and analysis of experimental designs for ratings. Cooperative Research Project No. 789, U. S. Office of Education, Washington, D. C., 1962.

Suydam, M. N. An instrument for evaluating experimental educational research reports. Journal of Educational Research, 1968, 61, 200-203.

Wandt, E. A cross-section of educational research. New York: David McKay Company, Inc., 1965.

Ward, A. W., Hall, B. W., & Schramm, C. F. Evaluation of published educational research: a national survey. American Educational Research Journal, 1975, 12, 109-128.

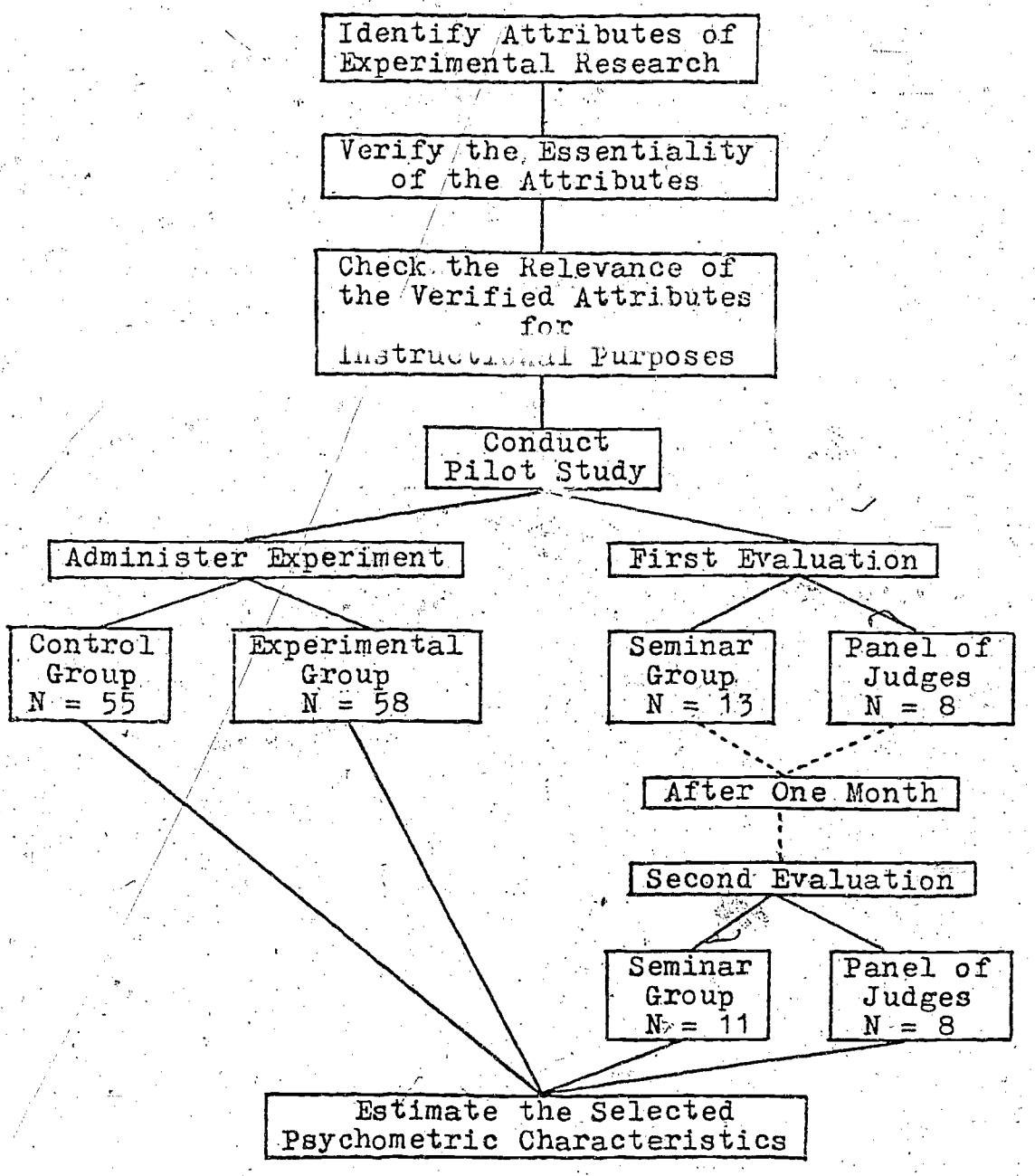

Figure 1. Flow Chart of Development Procedures of the Evaluation Instrument for Experimental Research (EIFER).

	Article 1		Article 2	
Rater	Pretest	Posttest	Pretest	Posttest
	1 . . . n	1 . . . n	1 . . . n	1 . . . n
1	Y_{1111}	. . . . .	. . . . .	1 . . . Y_{221n}
2	Y_{1121}	. . . . .	. . . . .	. . . . Y_{222n}
3	Y_{1131}	. . . . .	. . . . .	. . . . Y_{223n}
4	Y_{1141}	. . . . .	. . . . .	. . . . Y_{224n}
5	Y_{1151}	. . . . .	. . . . .	. . . . Y_{225n}
6	Y_{1161}	. . . . .	. . . . .	. . . . Y_{226n}
7	Y_{1171}	. . . . .	. . . . .	. . . . Y_{227n}
8	Y_{1181}	. . . . .	. . . . .	. . . . Y_{228n}

Figure 2. Analysis of Unreplicated A x B x C x D 4-Way Factorical Design, with 2 Articles, 8 Raters, Variable Traits, and 2 Occasions.

TABLE 1

KR-20 Reliability Coefficients for EIFER Section

Group:	Control[a]		Exp.[b]		Res. Seminar[c]				Judges[d]			
Article:	1		1		1		2		1		2	
Occasion:	Pre		Pre		Pre	Post	Pre	Post	Pre	Post	Pre	Post
Section												
Research Problem (13 items)	.71		.89	.94	.88	.87	.91	.95	.89	.82		
Related Literature (6 items)	.45		.76	.92	.82	.76	.63	.72	.82	.72		
Research Design (30 items)	.87		.94	.96	.91	.95	.88	.96	.76	.93		
Data Collection and Analysis (11 items)	.84		.78	.89	.83	.95	.80	.76	.65	.72		
Conclusions and Generalizations (10 items)	.68		.81	.94	.90	.91	.80	.94	.63	.31		
Style and Organization (3 items)	.72		.90	.94	.65	.83	.47	.65	.65	.52		
Total (73 items)	.93		.97	.99	.96	.98	.95	.98	.90	.91		
Global Evaluations (6 items)	.45		.64	.81	.97	.88	.93	.84	.90	.75	.76	

[a] $\underline{n}$ = 55

[b] $\underline{n}$ = 58

[c] $\underline{n}$ = 13 in pretest; 11 in posttest.

[d] $\underline{n}$ = 8.

TABLE 2

Matrix of Correlations Between EIFER Sections

	Test	1	2	3	4	5	6
1	Research Problem						
2	Related Literature	.55					
3	Research Design	.48	.53				
4	Data Collection and Analysis	.56	.55	.69			
5	Conclusions and Generalizations	.28	.47	.54	.55		
6	Style and Organization	.40	.30	.41	.50	.39	
7	Total*	.56	.63	.72	.78	.58	.50

*Corrected for overlap.

TABLE 3

Matrix of Item-Section Correlations[a]
(N = 58)

Item	Section[b] I	II	III	IV	V	VI	Total
1	.42*	.27	.37	.33	.30	.38	.44
2	.56*	.34	.41	.52	.31	.35	.54
3	.36*	.13	.14	.22	.23	.22	.25
4	.37	.10	.26	.36	.39	.42	.38
5	.27	.32	.23	.30	.21	.23	.32
6	.27	.32	.11	.20	.02	.16	.20
7	-.02	.08	.08	-.06	-.22	-.07	-.01
8	.20	.33	.23	.18	.24	.05	.27
9	.62*	.35	.39	.59	.22	.40	.54
10	.35	.39	.24	.36	.01	.09	.32
11	.50*	.48	.24	.19	.15	.15	.34
12	.28	.12	.30	.27	.19	.18	.32
13	.39*	.21	.13	.21	-.04	.16	.21
15	.19	.39*	.32	.28	.21	.03	.33
16	.32	.34	.27	.36	.51	.26	.41
17	.39	.12	.25	.45	.30	.31	.39
18	.35	.03	.32	.31	.26	.22	.37
19	.26	.08	.14	.12	-.01	.13	.16
20	.26	.41*	.36	.30	.26	.07	.38
22	.10	.39	.11	.19	.27	.27	.22
23	.50	.54	.44	.53	.32	.29	.57
24	.47	.39	.27	.44	.41	.41	.46
25	.25	.05	.31*	.16	.10	.14	.27
26	.28	.05	.33*	.16	-.01	.21	.28
27	.08	.15	.26	.30	.30	.18	.29
28	.45	.59	.35	.40	.51	.26	.51
29	.52	.39	.54*	.44	.28	.25	.57

Table 3—continued

Item	Section I	II	III	IV	V	VI	Total
30	.54	.47	.55*	.47	.46	.41	.70
31	.38	.16	.48*	.46	.11	.25	.47
32	.47	.40	.49	.63	.30	.35	.59
33	-.03	.04	.29*	.21	.09	-.04	.20
34	.04	.15	.25*	.13	.18	.16	.22
35	-.05	.04	.22	.15	.24	.14	.19
36	-.06	.09	.29*	.06	.11	-.02	.17
37	.18	.23	.41*	.25	.31	.03	.37
38	.45	.21	.59*	.35	.38	.30	.56
39	.13	.17	.46*	.34	.32	.19	.41
40	.13	.18	.41*	.25	.27	.30	.37
41	.23	.29	.16	.16	.09	.06	.21
42	.05	.24	.32*	.12	.03	-.06	.22
43	.45	.45	.49	.58	.44	.29	.60
44	.28	.40	.42*	.30	.34	.13	.44
45	.36	.29	.42	.44	.36	.34	.49
46	.16	.26	.48*	.27	.31	.23	.42
47	-.11	.23	.43*	.13	.05	-.04	.25
48	.10	.15	.51*	.37	.22	.19	.42
49	.23	.18	.54*	.45	.29	.34	.50
50	.12	.06	.54*	.35	.26	.10	.43
51	.16	.25	.50*	.30	.22	.14	.42
53	.32	.21	.41	.57*	.39	.37	.50
54	.38	.22	.59	.59*	.33	.45	.61
55	.39	.34	.50	.55*	.37	.33	.56
56	.21	.38	.23	.35	.28	.09	.32
57	.36	.44	.51	.63*	.43	.33	.60
58	.13	.19	.50	.43	.39	.14	.46
59	.52	.49	.61	.68*	.33	.45	.69

Table 3—continued

Item	Section I	II	III	IV	V	VI	Total
60	.50	.53	.55	.80*	.40	.49	.70
61	.65	.37	.34	.53	.24	.40	.52
62	.31	.35	.40	.42*	.25	.25	.45
63	.12	.29	.06	.20	.35	.19	.19
65	.25	.26	.09	.21	.08	.07	.18
66	.20	.36	.34	.40	.55*	.35	.44
67	.16	.30	.39	.37	.40*	.19	.41
68	.33	.48	.35	.35	.49*	.30	.46
69	.19	.15	.19	.07	.37*	.36	.25
70	.16	.23	.29	.18	.26	.13	.29
71	.13	.15	.40	.30	.33	.12	.37
72	-.01	.15	.16	.33	.38*	.28	.24
73	.01	.08	.31	.32	.33*	.25	.33
74	.08	.26	.19	.18	.19	-.06	.20
76	.32	.19	.34	.43	.30	.57*	.43
77	.46	.34	.35	.45	.34	.61*	.49
78	.20	.21	.29	.34	.30	.44*	.36

Note. Total items = 73 (excluding the 6 items for global evaluations).

[a]With df = 56, a correlation coefficient of .26 or above is significant at the .05 level.

[b]I: Research Problem; II: Related Literature; III: Research Design; IV: Data Collection and Analysis; V: Conclusions and Generalizations; VI: Style and Organization.

—The underlining identifies the EIFER section in which the item was placed.

*The item correlates higher with the section in which it is placed than it correlates with other sections.

TABLE 4

Articles x Raters x Traits x Occasions
Analysis of Variance:
EIFER Section One, Research Problem

Source	df	MS	F
Articles (A)	1	8.37	3.39
Occasions (D)	1	.41	.62
Raters (B)	7	1.23	.50
Traits (C)	12	.79	6.56**
A x D	1	.54	2.08
A x B	7	2.47	
D x B	7	.38	1.45
A x C	12	.09	.60
D x C	12	.22	1.91
B x C	84	.18	1.24
A x D x B	7	.26	
A x D x C	12	.11	1.59
A x B x C	84	.14	
D x B x C	84	.07	1.01
A x B x C x D	84	.07	

Note. Number of items = 13.

** $p < .01$.

TABLE 5

Articles x Raters x Traits x Occasions
Analysis of Variance:
EIFER Section Two, Related Literature

Source	df	MS	F
Articles (A)	1	7.52	10.23*
Occasions (D)	1	.08	1.93
Raters (B)	7	1.26	1.72
Traits (C)	5	.36	.76
A x D	1	.02	1.00
A x B	7	.74	
D x B	7	.04	1.71
A x C	5	.37	2.90*
D x C	5	.08	1.30
B x C	35	.23	1.80
A x D x B	7	.02	
A x D x C	5	.10	.77
A x B x C	35	.13	
D x B x C	35	.09	.75
A x B x C x D	35	.12	

Note. Number of items = 6.

*$p < .05$.

TABLE 6

Articles x Raters x Traits x Occasions
Analysis of Variance:
EIFER Section Three, Research Design

Source	df	MS	F
Articles (A)	1	5.25	3.53
Occasions (D)	1	.01	.01
Raters (B)	7	4.67	3.14
Traits (C)	29	1.37	5.35***
A x D	1	.30	6.86*
A x B	7	1.49	
D x B	7	.83	18.87***
A x C	29	.18	1.08
D x C	29	.20	1.88
B x C	203	.24	1.47**
A x D x B	7	.04	
A x D x C	29	.10	.96
A x B x C	203	.17	
D x B x C	203	.11	1.05
A x B x C x D	203	.10	

Note. Number of items = 30.

*$p < .05$.

**$p < .01$.

***$p < .001$.

TABLE 7

Articles x Raters x Traits x Occasions
Analysis of Variance:
EIFER Section Four, Data Collection and Analysis

Source	df	MS	F
Articles (A)	1	5.50	12.28**
Occasions (D)	1	.18	.30
Raters (B)	7	1.14	2.55
Traits (C)	10	.91	2.11
A x D	1	.92	1.87
A x B	7	.45	
D x B	7	.18	.37
A x C	10	.31	2.37*
D x C	10	.09	1.00
B x C	70	.25	1.88**
A x D x B	7	.49	
A x D x C	10	.07	.71
A x B x C	70	.13	
D x B x C	70	.12	1.24
A x B x C x D	70	.10	

Note. Number of items = 11.

*$p < .05$.

**$p < .01$.

TABLE 8

Articles x Raters x Traits x Occasions
Analysis of Variance:
EIFER Section Five, Conclusions and Generalizations

Source	df	MS	F
Articles (A)	1	3.00	2.72
Occasions (D)	1	.00	.01
Raters (B)	7	1.46	1.32
Traits (C)	9	1.02	5.20**
A x D	1	.25	1.82
A x B	7	1.10	
D x B	7	.30	2.18
A x C	9	.15	.97
D x C	9	.19	.92
B x C	63	.20	1.31
A x D x B	7	.14	
A x D x C	9	.22	1.61
A x B x C	63	.15	
D x B x C	63	.13	.92
A x B x C x D	63	.14	

Note. Number of items = 10.

**$p < .01$.

TABLE 9

Articles x Raters x Traits x Occasions
Analysis of Variance:
EIFER Section Six, Style and Organization

Source	df	MS	F
Articles (A)	1	2.67	6.22*
Occasions (D)	1	.67	2.80
Raters (B)	7	.55	1.28
Traits (C)	2	.59	3.59
A x D	1	.17	1.40
A x B	7	.43	
D x B	7	.19	1.60
A x C	2	.07	.86
D x C	2	.14	.32
B x C	14	.18	2.09
A x D x B	7	.12	
A x D x C	2	.45	2.98
A x B x C	14	.09	
D x B x C	14	.12	.82
A x B x C x D	14	.15	

Note. Number of items = 3.

*$p < .05$.

TABLE 10

Articles x Raters x Traits x Occasions
Analysis of Variance:
EIFER Section Seven, Total Test

Source	df	MS	F
Articles (A)	1	28.72	6.75*
Occasions (D)	1	.01	.03
Raters (B)	7	6.95	1.63
Traits (C)	72	1.11	3.62***
A x D	1	.02	.06
A x B	7	4.26	
D x B	7	.74	2.14
A x C	72	.23	1.31
D x C	72	.18	1.18
B x C	504	.25	1.47**
A x D x B	7	.35	
A x D x C	72	.15	1.36*
A x B x C	504	.17	
D x B x C	504	.11	1.07*
A x B x C x D	504	.11	

Note. Number of items = 73.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

TABLE 11

Estimated Average Correlation Between the Raters, Traits, and Occasions

Section	Interrater Reliability (r_{pp})	Intrarater Reliability (r_{II})	Stability of Traits (r_{TT})
Research Problem	.12	.60	.54
Related Literature	.33	.58	.44
Research Design	.06	.38	.18
Data Collection and Analysis	.21	.27	.33
Conclusions and Generalizations	.08	.30	.29
Style and Organization	.26	.28	.41
Total test	.14	.42	.26
Global Evaluations	.16	.55	.57

TABLE 12

Proportion of Responses to Each of the Six Overall Evaluation Items*

I. Research Problem

	0	1	
E	9 (.15)	49 (.35)	58
C	8 (.15)	47 (.85)	55
	17	96	

$s_{p_1 - p_2} = .067$, $z = 0$

II. Related Literature

	0	1	
E	13 (.22)	45 (.78)	58
C	13 (.24)	42 (.76)	55
	26	87	

$s_{p_1 - p_2} = .079$, $z = -.25$

III. Research Design

	0	1	
E	16 (.28)	42 (.72)	58
C	11 (.20)	44 (.80)	55
	27	86	

$s_{p_1 - p_2} = .08$, $z = 1$

IV. Data Collec. and Analy.

	0	1	
E	9 (.16)	49 (.84)	58
C	6 (.11)	49 (.89)	55
	15	98	

$s_{p_1 - p_2} = .063$, $z = -.79$

V. Conclu. and General.

	0	1	
E	13 (.22)	45 (.78)	58
C	9 (.16)	46 (.84)	55
	22	91	

$s_{p_1 - p_2} = .074$, $z = .81$

VI. Style and Organization

	0	1	
E	10 (.17)	48 (.83)	58
C	6 (.11)	49 (.89)	55
	16	97	

$s_{p_1 - p_2} = .065$, $z = .92$

*Each item corresponds to the section of the inventory.

- Appendix 1 -

Evaluation Instrument for Experimental Research (EIFER)

To Accompany

"An Inventory for Appraising Experimental Research Designed for Introductory Research Methods Classes"*

Yi-Ming Hsu
West Chester State College

Owen Scott
The University of Georgia

*Paper presented at the Annual Meeting of the American Educational Research Association, New York City, April, 1977.

EVALUATION OF RESEARCH REPORT[1]

Yi-Ming Hsu
West Chester State College

Instructions

The attached inventory is being formulated as a guide in evaluating reported experimental research in education and psychology. It was developed following an extensive review of literature to identify attributes in the research process which characterize experimental research. Each statement has been checked by nationally known experts in educational research methodology as essential to experimental research. In addition, in the judgment of professors of educational research across the country, students in the introductory course of research methods should be able to recognize and appraise the attributes as they are presented in published reports of experimental research.

The statements are organized under the following major headings to facilitate the rater in making evaluations:

I. RESEARCH PROBLEM
   A. Problem statement
   B. Hypothesis(es)

II. RELATED LITERATURE

III. RESEARCH DESIGN
   A. The population and sample
   B. The experimental arrangements
   C. Controls for the possible threats to the internal validity
   D. Controls for the possible threats to the external validity

IV. DATA COLLECTION AND ANALYSIS
   A. Data collection
   B. Data analysis

V. CONCLUSIONS AND GENERALIZATIONS

VI. STYLE AND ORGANIZATION OF THE REPORT

To make a proper use of the inventory, TWO different categories of evaluations are required:

(i) An evaluation of the specific statement of the attribute.
(ii) An overall or "global" evaluation of the aspect of the research process.

In judging the research with respect to each specific attribute, please use the following key:

1: The article contains no information concerning the attribute.
2: The information given clearly indicates that the attribute was inappropriately handled.
3: The information given suggests that the attribute may have been inproperly managed.
4: The information given indicates that the attribute was properly managed.
5: The information given clearly indicates that the attribute was appropriately handled.

The overall evaluation at the end of each section should NOT be determined by adding and/or averaging the responses to the separate statements. Instead, please make an overall or "global" appraisal of each aspect of the research using the same key and instructions (replacing the word "attribute" with "aspect") as you respond to the specific items.

The order in which the attributes are listed on the inventory will probably not be same as the order in which they are identified in the published reports of experimental research. For this reason you should follow the procedures described below so as to produce more dependable evaluations:

A. Read the statements on the inventory carefully.
B. Read the research report in its entirety without attempting to evaluate it.
C. Reread the report searching information relevant to the separate items on the inventory.
D. Refer to the key as often as needed in making your evaluation.
E. Mark the appropriate responses on the answer sheet provided (be sure to use #2 pencil).
F. Check the answer sheet to see that you have completed both each specific item evaluation and the SIX overall evaluations, and that you have written the information asked for.

EVALUATION INSTRUMENT FOR EXPERIMENTAL RESEARCH (EIFER)[2]

Yi-Ming Hsu
West Chester State College

I. RESEARCH PROBLEM

(i) Specific items

A. Problem statement

1. The research problem is clearly stated and precisely defined.
2. The significance of the study is demonstrated.
3. The relationship of the study to its scientific or experiential antecedents is indicated.
4. The objectives of the study are described.
5. Assumptions of the study are stated.
6. Limitations of the study are noted.
7. Critical or unusual terms are defined.

B. Hypothesis(es)

8. The hypothesis(es) is(are) easily identified.
9. The hypothesis(es) is(are) derived from the research problem.
10. The logical and empirical framework from which the hypothesis(es) was(were) derived is demonstrated.
11. The hypothesis(es) clearly identifies(fy) the independent variables.
12. The hypothesis(es) clearly identifies(fy) the effects to be measured by the dependent variable(s).
13. The hypothesis(es) is(are) testable.

(ii) Overall or "global" evaluation

14. RESEARCH PROBLEM

II. RELATED LITERATURE

(i) Specific items

15. Literature review is thorough and comprehensive.
16. Literature review is well-organized.
17. Literature reviewed is directly relevant to the research study.
18. The theoretical basis for the problem is identified.
19. The methodological strengths and/or weaknesses of the study are discussed.
20. The research design accounts for the variables which have probably influences on the dependent variable(s).

(ii) Overall or "global" evaluation

21. RELATED LITERATURE

III. RESEARCH DESIGN

(i) Specific items

A. The population and sample

22. The population to which generalization will be made is clearly specified.
23. The characteristics of the sample are fully described.
24. The sample size in the study is indicated.
25. Procedures for sample selection are fully described.
26. The method of assigning subjects to the comparison groups is clearly described.

B. The experimental arrangements

27. The treatment(s) is(are) randomly assigned to the comparison groups.
28. The research design includes the independent variables identified in the hypothesis(es).
29. The dependent variable(s) appropriately measures the effect(s) identified in the hypothesis(es).
30. The treatment(s) is(are) sufficiently described so that replication of the study may be possible.
31. Adequate information regarding the administration of the treatment(s) is provided.
32. The treatment(s) is(are) effectively applied in accordance with the objectives of the study.

C. The following possible threats to the internal validity of the experiment are controlled:

33. History: Specific events, external to the treatment conditions, occurring during the experimentation.
34. Maturation: Changes within the subject as a function of the passage of time during the course of experiment.
35. Testing: Variation between pretest and posttest responses due to cues from the pretest.
36. Instrumentation: Changes in the calibration of a measuring instrument or inconsistency of the scorers or raters can affect the measurements.
37. Statistical regression: Regression toward mean may occur if some but not all subjects are sampled from extreme groups.
38. Sample selection: Biases resulting from differences in the selection of subjects in the comparison groups.
39. Experimental mortality: The differential loss of subjects from the comparison groups during an experiment.
40. Interaction of selection and maturation, etc.: An interaction between selection and any other factors above which may be mistaken for the experimental effect.

D. The following possible threats to the external validity of the experiment are controlled:

(a) Population validity

41. Accessible vs. target population: Representativeness of the sample with respect to the population to which generalizations are made.
42. Interaction of personological variables and treatment effects: Reaction of the subjects with different personality characteristics to the treatment conditions.

(b) Ecological validity

43. Explicit definition of the independent variable: Descriptions of the management and operation of the treatment(s) (independent variables).
44. Multiple treatment interference: Interference with experimental results occurring from two or more treatments having been administered consecutively to the same subjects within a given time period.
45. Hawthorne effect: Awareness of the experiment may affect the response of the subject to the experimental stimuli.
46. Novelty and disruption effects: The experimental results may be due partly to the enthusiasm or disruption generated by the newness of the treatment.
47. Experimenter effect: Certain characteristics or behaviors of the experimenter may unintentionally influence the response of the subject.
48. Pretest sensitization: The administration of a pretest may have possible influences on the treatment effects.
49. Posttest sensitization: A test following the experiment may elicit effects which otherwise would remain latent or incomplete.
50. Interaction of history and treatment effects: The experimental results may be unique because of "extraneous" events occurring during the course of the experiment.
51. Interaction of time of measurement and treatment effects: Measurement of the dependent variable at two different points of time may produce two different results.

(ii) Overall or "global" evaluation

52. RESEARCH DESIGN

IV. DATA COLLECTION AND ANALYSIS

(i) Specific items

A. Data collection

53. The rationale for selection or development of the dependent variable measure(s) is clearly stated.
54. The measurement procedures adopted in the study for data gathering are specified.
55. Reliability data for the effects measurements are reported.
56. Validity data for the effects measurements are cited.
57. The procedures for data collection are carefully planned.
58. Deviations, if any, from that plan are made explicit.

B. Data analysis

59. The methods of data analysis are specifically described.
60. The methods of data analysis are appropriate for the specified research design.
61. The pattern of statistical analysis is applied correctly with respect to the nature of the raw data.
62. Statistical techniques are appropriate to the number of treatment groups and hypothesis(es) under consideration.
63. The level of significance in hypothesis testing is specified and adequate for the investigation.

(ii) Overall or "global" evaluation

64. DATA COLLECTION AND ANALYSIS

V. CONCLUSIONS AND GENERALIZATIONS

(i) Specific items

65. Tables and figures display the data basic to testing the hypothesis(es).
66. Results of hypothesis testing are reported with support of statistical evidence.
67. Claims for the probable truth or falsity of the research hypothesis(es) are supported by the evidence presented.
68. Conclusions are objectively stated and effectively summarized.
69. Discussions are consistent with the results presented.
70. The extent to which the study can be generalized to the population of interest is clearly identified.
71. Generalizations made are reasonable and logical.
72. Evidence is presented in support of the internal validity of the study.
73. The findings are related to the previous research on the problem of inquiry.
74. Problems raised from the study are stated for further exploration.

(ii) Overall or "global" evaluation

75. CONCLUSIONS AND GENERALIZATIONS

VI. STYLE AND ORGANIZATION OF THE REPORT

(i) Specific items

76. The report is written in clear, understandable language.
77. Organization of the content is clear and rigorous.
78. The style and tone of the report reflect an objective, unbiased, and scientific attitude.

(ii) Overall or "global" evaluation

79. STYLE AND ORGANIZATION OF THE REPORT